Design and Analysis of Group-Randomized Trials

David M. Murray, Ph.D.
Lillian and Morrie Moss Chair of Excellence
Department of Psychology
University of Memphis

Effective September 1, 2005
Chair, Division of Epidemiology and Biostatistics
School of Public Health
The Ohio State University

References

- Primary Reference
  - Murray, D.M. Design and Analysis of Group-Randomized Trials. New York: Oxford University Press, 1998.
  - http://www.oup-usa.org/isbn/0195120361.html
    http://www.psyc.memphis.edu/faculty/Murray/correct.htm
- Secondary References
  - Varnell, S., Murray, D. M., Janega, J. B., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent practices. American Journal of Public Health, 94(3), 393-399.
  - Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. American Journal of Public Health, 94(3), 423-432.

References

- Secondary References (cont.)
  - Blitstein, JL, Hannan, PJ, Murray, DM, & Shadish, WR. (2005). Increasing the degrees of freedom in existing group randomized trials through the use of external estimates of intraclass correlation: The DF* approach. Evaluation Review, 29(3), 241-267.
  - Blitstein, JL, Murray, DM, Hannan, PJ, & Shadish, WR. (2005). Increasing the degrees of freedom in future group randomized trials: The df* approach. Evaluation Review, 29(3), 268-286.
  - Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. Evaluation Review, 27(1), 79-103.

Examples

- School trials
  - Trial of Activity for Adolescent Girls (TAAG)
  - Childhood and Adolescent Trial for Cardiovascular Health (CATCH)
- Worksite trials
  - Working Well Trial (WWT)
  - SUCCESS
- Community trials
  - Rapid Early Action for Coronary Treatment (REACT)
  - Community Youth Development Study (CYDS)

Distinguishing Characteristics

- The unit of assignment is an identifiable group.
- Different groups are allocated to each condition.
- The units of observation are members of the groups.
- The number of groups allocated to each condition is usually limited.

Impact on the Design

- In any single realization of the experiment, there is limited opportunity for randomization to distribute all potential sources of bias evenly.
- As a result, bias is more of a concern in GRTs than in many RCTs.
- This increases the need to use design strategies that will limit bias.

Impact on the Analysis

- The members of the same group will share some physical, geographic, social or other connection.
- That connection will create a positive intraclass correlation that reflects an extra component of variance attributable to the group.

$$ICC_{m:g:c} = corr\left(y_{i:k:l}, y_{i':k:l}\right)$$

- The positive ICC reduces the variation among the members of the same group so the within-group variance is:

$$\sigma_e^2 = \sigma_y^2\left(1 - ICC_{m:g:c}\right)$$

Impact on the Analysis

- The between-group component is the one's complement:

$$\sigma_{g:c}^2 = \sigma_y^2\left(ICC_{m:g:c}\right)$$

- The total variance is the sum of the two components:

$$\sigma_y^2 = \sigma_e^2 + \sigma_{g:c}^2$$

- The intraclass correlation is the fraction of the total variation in the data that is attributable to the unit of assignment:

$$ICC_{m:g:c} = \frac{\sigma_{g:c}^2}{\sigma_e^2 + \sigma_{g:c}^2}$$

Impact on the Analysis

- Given m members in each of g groups...
  - When group membership is established by random assignment, $\sigma_{\bar{y}_g}^2 = \frac{\sigma_y^2}{m}$
  - When group membership is not established by random assignment, $\sigma_{\bar{y}_g}^2 = \frac{\sigma_e^2}{m} + \sigma_g^2$
  - Or equivalently, $\sigma_{\bar{y}_g}^2 = \frac{\sigma_y^2}{m}\left(1 + (m-1)\,ICC\right)$

Impact on the Analysis

- The variance of any group-level statistic will be larger when identifiable groups are assigned to conditions.
- With a limited number of groups in each condition, the df to estimate the group-level component of variance will be limited.
- Any analysis that ignores the extra variation or the limited df will have a Type I error rate that is inflated, often badly.
- Extra variation and limited df serve to limit power, so they must be considered at the design stage to ensure adequate power.

The Warning

Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception, however, and should be discouraged.

Cornfield (1978, p. 101-102)

Questions About the Value of GRTs

- Disappointing results for several large trials in the early 1990s led some to question the value of group-randomized trials.
- To question group-randomized trials in general based on these results is short sighted and impractical.
- A group-randomized trial remains the best design available whenever the investigator wants to evaluate an intervention that…
  - operates at a group level
  - manipulates the social or physical environment
  - cannot be delivered to individuals
- The research community understands this, and the number of GRTs published in good journals has doubled in the last ten years.

The Challenge

...we should not abandon community trials but should gather the knowledge necessary to refine them.

Susser (1995, p. 158)

- The challenge is to create trials that are:
  - Rigorous enough to avoid threats to validity of the design,
  - Analyzed so as to avoid threats to statistical validity,
  - Powerful enough to provide an answer to the question,
  - And inexpensive enough to be practical.
- The question is not whether to conduct GRTs, but rather how to do them well.

Planning the Trial

- The driving force behind any study must be the research question.
  - The question will identify the target population, the setting, the endpoints, and the intervention.
  - Those factors will shape the design and analysis plan.
- The primary criteria for choosing that question should be:
  - Is it important enough to do?
    - Will the trial address an important public health question?
    - Will the results advance the field?
  - Is this the right time to do it?
    - Is there preliminary evidence of feasibility and efficacy for the intervention?
    - Are there good estimates for the parameters required to size the study?
- The investigators should keep the question in mind.

Fundamentals of Research Design

- The goal in any comparative trial is to allow valid inference that the intervention as implemented caused the result as observed.
- Three elements are required:
  - Control observations
  - A minimum of bias in the estimate of the intervention effect
  - Sufficient precision for that estimate
- The three most important tools to limit bias and improve precision in any comparative trial, including a GRT, are:
  - Randomization
  - Replication
  - Variance reduction

Potential Threats to Internal Validity

- Four primary threats:
  - Selection
  - History and differential history
  - Maturation and differential maturation
  - Contamination

Strategies to Limit Threats to Internal Validity

- Randomization
- A priori matching or stratification of groups
- Objective measures
- Independent evaluation personnel who are blind to conditions
- Analytic strategies
- Avoid the pitfalls that invite threats to internal validity
  - Testing and differential testing
  - Instrumentation and differential instrumentation
  - Regression to the mean and differential regression to the mean
  - Attrition and differential attrition

Threats to the Validity of the Analysis

- Misspecification of the analysis model
  - Ignore a measurable source of random variation
  - Misrepresent a measurable source of random variation
  - Misrepresent the pattern of over-time correlation in the data
- Low power
  - Weak interventions
  - Insufficient replication of groups and time intervals
  - High variance or intraclass correlation in endpoints
  - Poor reliability of intervention implementation

Strategies to Protect the Validity of the Analysis

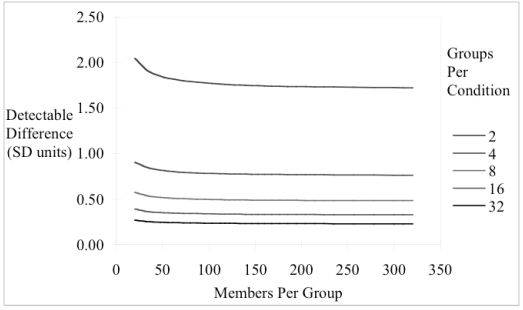
- Avoid model misspecification
  - Plan the analysis concurrent with the design.
  - Plan the analysis around the primary endpoints.
  - Anticipate all sources of random variation.
  - Anticipate patterns of over-time correlation.
  - Consider alternate models for time.
  - Assess potential confounding and effect modification.

Strategies to Protect the Validity of the Analysis

- Avoid low power
  - Employ strong interventions with good reach.
  - Maintain reliability of intervention implementation.
  - Employ more and smaller groups instead of a few large groups.
  - Employ more and smaller surveys or continuous surveillance instead of a few large surveys.
  - Employ regression adjustment for covariates to reduce variance and intraclass correlation.
  - Consider matching or stratification in the analysis.

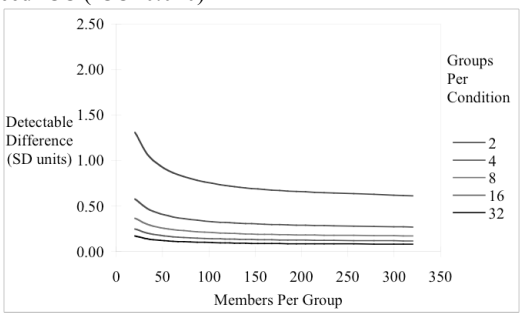
Strategies to Improve Precision

- Increased replication (ICC=0.100)



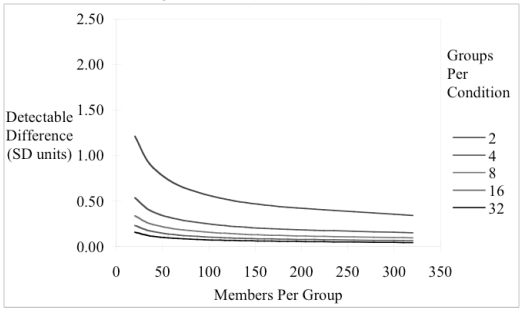
Strategies to Improve Precision

- Reduced ICC (ICC=0.010)



Strategies to Improve Precision

- The law of diminishing returns (ICC=0.001)



Power for Group-Randomized Trials

- The usual methods must be adapted to reflect nested design
  - The variance is greater in a GRT due to the expected ICC.
  - Df should be based on the number of groups, not members.
- A good source on power in GRTs is Chapter 9 in Murray (1998).
- Many papers now report ICCs and show how to plan a GRT.
  - For lists of papers, cf. Murray & Blitstein, 2003 and Murray et al., 2004.
- Two mew papers show how to combine ICC estimates from multiple sources to get more df than would be available otherwise.
  - cf. Blitstein et al., 2005a, b.
- Power in GRTs is tricky, and investigators are advised to get help from someone familiar with these methods.

A Classification Scheme for Statistical Models

	Gaussian Distribution	Non-Gaussian Distribution
One Random Effect	General Linear Model	Generalized Linear Model
Two Or More Random Effects	General Linear Mixed Model	Generalized Linear Mixed Model

Fig 4.1

- Fixed effect: the investigators want to draw inferences about the levels used in the study.
- Random effect: the investigators want to draw inferences about some larger population of levels that are only represented by the levels used in the study.

Preferred Analytic Strategies for Designs Having One or Two Time Intervals

- Mixed-model ANOVA/ANCOVA
  - Extension of the familiar ANOVA/ANCOVA based on the General Linear Model.
  - Fit using the General Linear Mixed Model or the Generalized Linear Mixed Model.
  - Accommodates regression adjustment for covariates.
  - Can not misrepresent over-time correlation.
  - Can take several forms
    - Posttest-only ANOVA/ANCOVA
    - ANCOVA of posttest with regression adjustment for pretest
    - Repeated measures ANOVA/ANCOVA for pretest-posttest design
  - Simulations have shown that these methods have the nominal Type I error rate across a wide range of conditions common in GRTs.

Preferred Analytic Strategies for Designs Having More Than Two Time Intervals

- Random coefficients models
  - Mixed-model ANOVA/ANCOVA assumes homogeneity of group-specific slopes.
  - Simulations have shown that mixed-model ANOVA has an inflated Type I error rate if those slopes are heterogeneous.
  - Random coefficients models allow for heterogeneity of those slopes.
  - Random coefficients models have the nominal Type I error rate across a wide range of conditions common in GRTs.
  - Random coefficients models are used increasingly in the evaluation of public health interventions.
    - Examples include NCI's Project ASSIST and NHLBI's REACT.

What About Randomization Tests?

- The intervention effect is a function of unadjusted or adjusted group-specific means, slopes or other group-level statistic.
- Under the null hypothesis of no intervention effect, the actual arrangement of those group-level statistics among the study conditions is but one of many equally likely arrangements.
- The randomization test systematically computes the effect for all possible arrangements.
- The probability of getting a result more extreme than that observed is the proportion of effects that are greater than that observed.
- No distributional or other assumptions are required.

What About Randomization Tests?

- Strengths
  - Gail et al. (1996) reported that randomization tests had nominal Type I and II error rates across a variety of conditions common to GRTs.
  - Randomization does ensure the nominal Type I error rate, even when very few heterogeneous groups are assigned to each condition.
  - Programs for randomization tests are available in print and on the web.
- Weaknesses
  - The unadjusted randomization test does not offer any more protection against confounding than other unadjusted tests.
  - Regression adjustment for covariates requires many of the same assumptions as the model-based tests.
  - Randomization tests provide only a point estimate and a p-value, where model-based methods provide parameter estimates, standard errors, etc.

What About a Method Like GEE That is Robust Against Misspecification?

- Methods based on GEE use an empirical sandwich estimator for standard errors.
- That estimator is asymptotically robust against misspecification of the random-effects covariance matrix.
- When the degrees of freedom are limited (<40), the empirical sandwich estimator has an unpredictable Type I error rate.
- Recent work provides corrections for that problem, but they are not yet incorporated into the standard software.
- Methods that employ the corrected empirical sandwich estimator may have broad application in GRTs.

What About Fixed-Effect Methods in Two Stages?

- Introduced as the first solution to the unit of analysis problem in the 1950s.
- Commonly known as the means analysis.
- Simple to do and easy to explain.
- Gives results identical to the mixed-model ANOVA/ANCOVA if both are properly implemented.
- Can be adapted to perform random coefficients analyses.
- Can be adapted to complex designs where one-stage analyses are not possible.
- Used in several large trials, including CATCH, MHHP, and REACT.

What About a Post Hoc Correction to the Usual Fixed-Effects Methods?

- Several have proposed a post hoc correction of the fixed-effect F-test by dividing it by Kish's DEFF estimated from other data.
- This would rely on the strong assumption that the external estimate is valid for the data at hand.
- There is no consensus on the df to be used to evaluate the corrected test statistic.
- Recent work has offered guidelines for selecting and combining estimates and using the precision of the combined estimate to calculate df for the adjusted test statistic.
  - cf. Blitstein et al. 2005a, b

What About Methods Developed for Analysis of Complex Survey Samples?

- Methods developed for analysis of complex survey samples perform well given a large number of primary sampling units.
- These methods do not perform well when the number of primary sampling units is limited (<40).
- The standard normal approximation that often accompanies these methods is not appropriate given limited df.
- Those methods for analysis of complex survey samples may have limited application in GRTs.
- Many survey analysis programs have adopted empirical sandwich estimation, and if one of the small-sample correction factors is employed, such methods would be applicable to GRTs.

What About Analysis by Subgroups?

- Some have suggested analysis by subgroup rather than group, especially when the number of groups is limited.
  - Classrooms instead of schools
  - Physicians instead of clinics
- This approach rests on the strong assumption that the subgroup captures all of the variation due to the group.
- This approach has an inflated Type I error rate even when the subgroup captures 80% of the group variation.

- Analysis by subgroups instead of groups is not recommended.

What About Deleting the Unit of Assignment From the Model if it is not Significant?

- The df for such tests are usually limited; as such, their power is usually limited.
- Standard errors for variance components are not well estimated when the variance components are near zero.
- Even a small ICC, if ignored, can inflate the Type I error rate if the number of members per group is moderate to large.
- The prudent course is to retain all random effects associated with the study design and sampling plan.

What About Studies Based on Only One Group per Condition?

- Cannot separately estimate variation due to the group and variation due to condition.
- Must rely on a strong assumption:
  - Post hoc correction: external estimate is valid
  - Subgroup or batch analysis: subgroup captures group variance
  - Fixed-effects analysis: group variance is zero
- Varnell et al. (2001) found the second and third strategies are likely to have an inflated Type I error rate.
  - This design should be avoided if causal inference is important.
  - It may still be helpful for preliminary studies.

What About Studies That Randomize Individuals but Deliver Treatments to Groups?

- Many studies randomize participants as individuals but deliver treatments in small groups.
  - Psychotherapy, weight loss, smoking cessation, etc.
  - Little or no group-level ICC at baseline.
  - Positive ICC later, with the magnitude proportional to the intensity and duration of the interaction among the group members.
- Analyses that ignore the ICC risk an inflated Type I error rate.
  - Not as severe as in a GRT, but can exceed 15% under conditions common to these studies.
  - The solution is the same as in a GRT

Is the Non-Negativity Constraint OK?

- Software based on maximum likelihood routinely constrains variance estimates to be non-negative.
- Simulation studies published in 1984 indicated that this constraint introduced a positive bias in the estimates of the variance components.
- More recent simulations for GRTs showed that the constraint depressed the Type I error rate, often dramatically.
- Analysts should avoid the non-negativity constraint.

State of the Science on Methods

- GRTs require analyses that reflect the nested designs inherent in these studies.
- Used alone, the usual methods based on the General or Generalized Linear Model are not valid.
- Methods based on the General Linear Mixed Model and on the Generalized Linear Mixed Model are widely applicable.
  - For designs having one or two time intervals, mixed-model ANOVA/ANCOVA is recommended.
  - For designs having three or more time intervals, random coefficients models are recommended.
- Other methods can be used effectively, with proper care, including randomization tests, GEE and two-stage methods.

State of the Practice

- The first review of GRTs was published by Donner et al. in 1990.
  - Only 19% took the ICC into account in their sample size calculations.
  - Only 50% took the ICC into account in their analysis.
- A review by Simpson et al. in 1995 reported little progress.
  - Only 19% took the ICC into account in their sample size calculations.
  - Only 57% took the ICC into account in their analysis.
- We set out to replicate the review by Simpson et al., who had considered all GRTs in Preventive Medicine and AJPH, 1990-93.
- We reviewed all GRTs published in the same journals, 1998-02.
  - Varnell, S., Murray, D. M., Janega, J. B., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent practices. *American Journal of Public Health, 94*(3), 393-399.

Procedures

- Studies had to use randomization to assign identifiable social groups to study conditions, with observations taken on members of those groups to assess the impact of an intervention.
- Where the paper referred to an earlier "design paper", we also reviewed that paper.
- Each reviewer independently completed a review form, assessing the article on a variety of items related to design, sample size estimation, and analysis.
- The reviewers discussed each paper as a group and any disagreements were resolved in discussion.

Findings

- 58 studies met the inclusion criteria.
- 27 background "design" papers.
- 47% of papers in AJPH
- 53% of papers in Preventive Medicine
- 11.6 GRT papers per year, vs. 5.3 per year in Simpson et al.

Table 2- Characteristics of Studies Described in the 58 Reviewed Articles

Characteristic	Number of Articles	%
Number of Study Conditions		
Two	49	84.5
Three	6	10.3
Four or more	3	5.2
Matching or Stratification in Design		
Matching	20	34.5
Stratification	18	31.0
Matching and Stratification	7	12.1
Randomization without Matching or Stratification	13	22.4
Type of Group		
Schools or Colleges	17	29.3
Worksites	11	19.0
Medical Practices	9	15.5
Communities, Neighborhoods or Postal Networks	7	12.1
Housing Projects or Apartment Buildings	3	5.2
Churches	3	5.2
Other	8	13.8

Table 2- Characteristics of Studies Described in the 58 Reviewed Articles

Characteristic	Number of Articles	%
Number of Groups per Condition		
1 Group	3	5.2
2-3 Groups	5	8.6
4-5 Groups	7	12.1
6-12 Groups	18	31.0
13-25 Groups	18	31.0
> 25 Groups	7	12.1
Number of Members per Group		
<10 Members	8	13.8
10-50 Members	19	32.8
51-100 Members	14	24.1
>100 Members	17	29.3
Number of Timepoints		
1 Timepoint	2	3.4
2 Timepoints	32	55.2
3 Timepoints	17	29.3
4-9 Timepoints	6	10.3
Number of Timepoints Varies Within Study	1	1.7

Table 2- Characteristics of Studies Described in the 58 Reviewed Articles

Characteristic	Number of Articles	%
Design		
Cohort	37	63.8
Cross-sectional	12	20.7
Combination of Cohort and Cross-sectional	9	15.5
Primary Outcome Variables		
Smoking Prevention or Cessation	15	25.9
Dietary Variables	12	20.7
Health Screening	7	12.1
Alcohol, Drug, or Combination of Alcohol, Tobacco, Drugs	5	8.6
Multiple Health Measures	5	8.6
Sun Protection	3	5.2
Preventing Physical or Sexual Abuse	2	3.4
Physician Preventive Practices	2	3.4
Workplace Health and Safety Measures	2	3.4

Table 3- Results of the review of analytic methods

Criteria	N(%) (Total=57)
Articles reporting only appropriate methods	31 (54.4%)
Method	
Mixed-model methods with baseline measurement as covariate	10 (17.5%)
Mixed-model ANOVA/ANCOVA approach with one or two timepoints	9 (15.8%)
Generalized Estimating Equations with 40 or more groups	2 (3.5%)
Two-stage analysis (analysis of group means or other summary statistics)	11 (19.3)
Articles reporting only inappropriate methods	11 (19.3%)
Method	
Analysis at an individual level, ignoring group-level ICC	6 (10.5%)
Analysis at a subgroup level, ignoring group-level ICC	3 (5.3%)
Analysis with group as a fixed effect	1 (1.8%)
Mixed-model ANOVA/ANCOVA approach with more than two timepoints	1 (1.8%)
Generalized Estimating Equations with fewer than 40 groups	4 (7.0%)

Table 3- Results of the review of analytic methods

Criteria	N(%) (Total=57)
Articles reporting some appropriate and some inappropriate methods	15 (26.3%)
Appropriate Methods	
Mixed-model methods with baseline measurement as covariate	6 (10.5%)
Mixed-model methods with one or two timepoints	4 (7.0%)
Generalized Estimating Equations with 40 or more groups	2 (3.5%)
Two-stage analysis	3 (5.3%)
Inappropriate Methods	
Analysis at an individual level, ignoring group-level ICC	9 (15.8%)
Analysis at a subgroup level, ignoring group-level ICC	1 (1.8%)
GEE or other asymptotically robust method with fewer than 40 groups	3 (5.3%)

Discussion

- It has been 27 years since Cornfield drew attention in the public health literature to the special design and analytic issues in GRTs.
- Since then, a very large literature has developed on appropriate design and analytic methods.
- Two textbooks now offer comprehensive treatments.
- Hundreds of other articles have appeared, along with many related books.
- Readily available software supports appropriate analytic methods.

Discussion

- Only 15.5% of the articles reported evidence of using appropriate methods for sample size calculations.
  - 46.6% of the articles had fewer than 10 groups per condition.
  - Of those, only 1 reported evidence of appropriate sample size calculations.
- Only 54.4% of the articles reported only analyses judged to be appropriate.
- Fully 19.3% reported only analyses deemed invalid by the earlier reviews.
- The remaining 45.6% reported a mix of appropriate and inappropriate analyses.

Discussion

- Some investigators, reviewers and editors still have not heard or accepted the long-standing warnings against analyses that…
  - Ignore the group-level ICC altogether
  - Include group as a fixed effect
- Many have not heard about the more recent warnings that other methods are inappropriate under certain conditions…
  - Standard GEE when there are fewer than 40 groups in the study.
  - Repeated measures ANOVA/ANCOVA models with > two time points.
- We still see designs with one group randomized to each condition.
  - They have been described as having "no valid analysis" absent strong and untestable assumptions.

Recommendations

- Reviewers and editors need to be more vigilant.
  - Journals should require review by a methodologist familiar with GRT issues for all GRT articles submitted.
- Authors need to rely more heavily on well trained methodologists.
- We need to repeat this kind of review in other journals and periodically over time to see whether the field is making progress.

Planning Future Group-Randomized Trials

- Address an important research question.
- Employ an intervention that has a strong theoretical base and preliminary evidence of feasibility and efficacy.
- Randomize enough assignment units to have good power.
- Design the trial in recognition of the major threats to the validity of the design and analysis of GRTs.
- Employ good quality-control measures.
- Employ good process-evaluation measures.
- Employ reliable and valid measures of the primary endpoints.
- Analyze using methods appropriate to the design and the structure of the primary endpoints.
- Interpret carefully.